

SOARAN: A Service-oriented Architecture for Radio Access Network Sharing in Evolving Mobile Networks

Jun He, and Wei Song, *Member, IEEE*

Faculty of Computer Science

University of New Brunswick, Fredericton, Canada

Emails: {jhe2, wsong}@unb.ca

Abstract

Mobile networks are undergoing fast evolution to software-defined networking (SDN) infrastructure in order to accommodate the ever-growing mobile traffic and overcome the network management nightmares caused by unremitting acceleration in technology innovations and evolution of the service market. Enabled by virtualized network functionalities, evolving carrier wireless networks tend to share radio access network (RAN) among multiple (virtual) network operators so as to increase network capacity and reduce expenses. However, existing RAN sharing models are operator-oriented, which expose extensive resource details, *e.g.*, infrastructure and spectrum, to participating network operators for resource-sharing purposes. These old-fashioned models violate the design principles of SDN abstraction and are infeasible to manage the thriving traffic of on-demand customized services. This paper presents SOARAN, a service-oriented framework for RAN sharing in mobile networks evolving from LTE/LTE advanced to software-defined carrier wireless networks (SD-CWNs), which decouples network operators from radio resource by providing application-level differentiated services. SOARAN defines a series of abstract applications with distinct Quality of Experience (QoE) requirements. The central controller periodically computes application-level resource allocation for each radio element with respect to runtime traffic demands and channel conditions, and disseminates these allocation decisions as service-oriented policies to respect element. The radio elements then independently determine flow-level resource allocation within each application to accomplish these policies. We formulate the application-level resource allocation as an optimization problem and develop a fast algorithm to solve it with a provably approximate guarantee. The efficacy of SOARAN is validated through theoretical analysis and computer simulations. We also show that SOARAN is in line with the design of SD-CWNs.

Index Terms

Radio access network, RAN sharing, software-defined RAN, resource virtualization, network abstraction

I. INTRODUCTION

Nowadays, the mobile system has become a service-oriented platform swamped with millions of applications providing differential services to data-hungry devices whose owners increasingly regard the ubiquitous quality-guaranteed network services as a human right, regardless of traffic overburdening of the network and high costs of system upgrading. Sharing radio access networks (RANs) among mobile network operators (MNOs) and mobile virtual network operators (MVNOs) is not only a promising way to expand system capacity and reduce both capital expenses (CAPEX) and operational expenses (OPEX) [1], but also an inevitable trend of carrier wireless networks facing the harsh realities of unremitting growth of traffic demand and continuously declining unit-data revenue. MVNOs are a set of service providers (SPs) or network resellers, *e.g.*, content providers (CPs), who do not own network infrastructure but share MNOs' RANs in a rental manner based on service level agreements (SLAs) with them. Along with the proliferation of smart devices, MVNOs play an essential role in enriching mobile networks with innovative applications, differentiated services, and prompting subscribers' engagements [2], shifting mobile networks from operator-oriented systems to service-oriented systems. However, despite the rapid development in software defined carrier wireless networks (SD-CWNs), service-oriented RAN sharing remains unexplored.

The oldest form of RAN sharing in MNOs is to accommodate "foreign" subscribers roaming from other networks. The Long-Term Evolution (LTE) system of the 3rd Generation Partnership Project (3GPP) supports RAN sharing among core network operators in either a gateway core network (GWCN) configuration or a multi-operator core network (MOCN) configuration with shared or independent mobility management entities (MMEs), respectively. Fig. 1 shows the network model of sharing RAN we study in this paper. MNOs share the RAN through a radio access gateway, *e.g.*, the serving gateway in LTE or the access service network (ASN) gateway in WiMAX, and provide RAN access to MVNOs via their IP cores. Radio elements refer to base-stations, *e.g.*, eNodeB, pico/micro cells, that are managed by a centralized controller at the RAN gateway and provide radio access to subscribers.

As mobile networks are merging into the cloud, RANs are also undergoing fast evolution to SD-CWNs with virtualized radio resource [3]. With the promotion of scalability and manageability, virtualized RANs have developed maturing methods for resource slicing and frame scheduling, which eases the resource

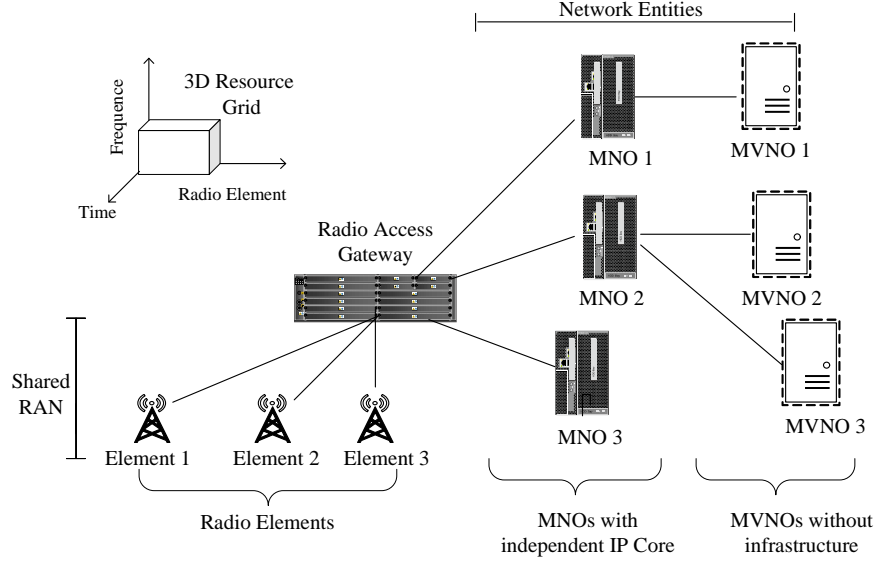


Fig. 1. RAN Sharing Network Model.

management at the controller [4,5]. As depicted in Fig. 1, the radio resource over the RAN is abstracted in a configurable *3-dimensional resource grid* of radio element index, frequency and time [4]. The central controller has a view of one virtual “big” base-station upon which radio resource is slicable and allocatable via the northbound application programming interfaces (APIs) that the virtualized RAN provides to the controller.

State-of-the-art resource schedulers, *e.g.*, [5,6], divide resource among operators sharing the RAN based on SLAs with operator isolation in a resource-reservation manner. That is, SLAs specify the resource shares of each operator either on a per-base-station basis (*e.g.*, [5]) or on a RAN basis (*e.g.*, [6]). For instance, in a network with 2 operators, operator 1 reserves 30% of the resource (overall or per-base-station) while operator 2 takes 70%. These share ratios could also be a range, *e.g.*, minimal 20% and maximal 35%, to enable adaptive resource scheduling according to data traffic [6]. The allocation decisions made at the controller are then applied by lower-layer frame schedulers. However, due to the following concerns, we argue that these operator-oriented designs are against the principles of network virtualization and will soon become infeasible in the expanding mobile networks.

- These methods expose extensive details of RAN to the operators, *e.g.*, the number, distribution and capacity of radio elements. Even though the resource share is given in percentages, for pricing and budget purposes, one operator needs to know the coverage of the RAN and the bandwidth it provides.

Therefore, these operator-oriented approaches will make the network management and RAN upgrade even more complicated as the RAN or the number of operators grows.

- To fulfill SLAs, the controller requires the ownership information of each data flow, which can be retrieved in no way but by conducting deep packet inspections (DPIs). Apparently, the overhead of DPI per flow would be intolerable.
- It is hard to manage Quality of Experience (QoE) in these models. Each operator provides a set of services (regarded as applications hereafter) with differentiated QoE requirements. Since the resource is allocated at the operator level, each operator independently needs to manage QoE of its own applications, *e.g.*, using bearers [7], which not only plunges QoE support at the RAN gateway into chaos, but also produces an aggregate resource utility arbitrarily suboptimal.
- These existing works determine detailed resource allocation (*e.g.*, per flow) at the central controller, which is incompatible to new developments in RAN with heterogeneous radio elements due to the overwhelming overhead of reporting wireless channel details. For example, only symmetric base-stations are considered in [6].

To address these problems, we proposed an application sepecified RAN sharing architecture in [8]. In this paper, we extend the intention of [8] and re-define the RAN sharing model, namely SOARAN, a service-oriented architecture for RAN sharing in mobile networks. Instead of promising certain amount of resource for each operator, SOARAN defines a configurable set of abstract applications with respect to differentiated QoE requirements such that operators can map their concrete applications to the abstract ones and determine application-specified bandwidth they need. Operators can now focus on their application-level demands through an application-abstraction layer provided by SOARAN, while SOARAN takes care of lower-layer resource allocation. In this way, SOARAN decouples operators from radio resource allocation, keeps the upgrade of RAN facilities and resource allocation transparent to operators and enables better resource virtualization.

A new model of SLAs is defined based on the promised services for each operator. The charging policy is then determined by the service package each operator purchases, *e.g.*, in the form of a series of (application, bandwidth) tuples. Such service-oriented SLA model makes a step towards merging RANs into the cloud. SOARAN also develops an optimization framework and a fast algorithm to optimize the

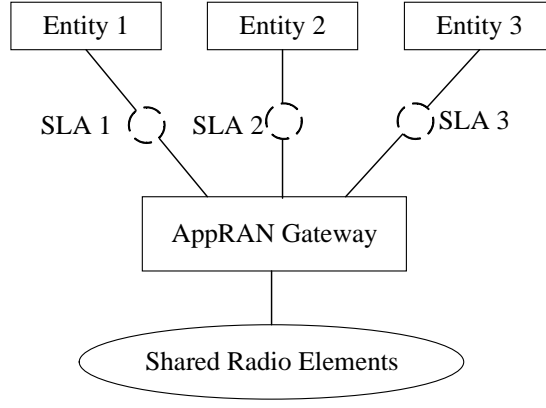


Fig. 2. SOARAN Network Model.

resource allocation. The optimization framework takes average resource-to-bandwidth conversion ratios reported by radio elements as input and computes the optimal resource allocation among applications for each radio element. With negotiable overhead, the central controller determines the optimal resource allocation on the application level.

The rest of paper is organized as follows. Section II describes the design of SOARAN. The kernel resource allocation algorithm is presented in Section III. Section IV provides the numerical results and Section V concludes the paper.

II. DESIGN OF SOARAN

SOARAN flattens the RAN sharing structure and re-models it in Fig. 2. Network operators including MNOs, MVNOs, are regarded as *entities* driving data flows to the SOARAN gateway, where the flows are differentially treated according to SLA configurations. An MNO with several virtual operators attached to it (see Fig. 1) can be treated as one entity or several entities as it describes, while the details inside are kept transparent to SOARAN, granting more flexibility to the MNO.

A. The SOARAN Model

SOARAN defines a series of *abstract applications* with respect to differentiated QoE levels, which can be readily supported using RAN “bearers” in 3GPP systems [7]. Fig. 3 shows an example of the abstract application table. The description of an abstract application consists of identification information (id and priority), QoE guarantees to entities (delay, packet loss rate, *etc.*), action information to radio elements

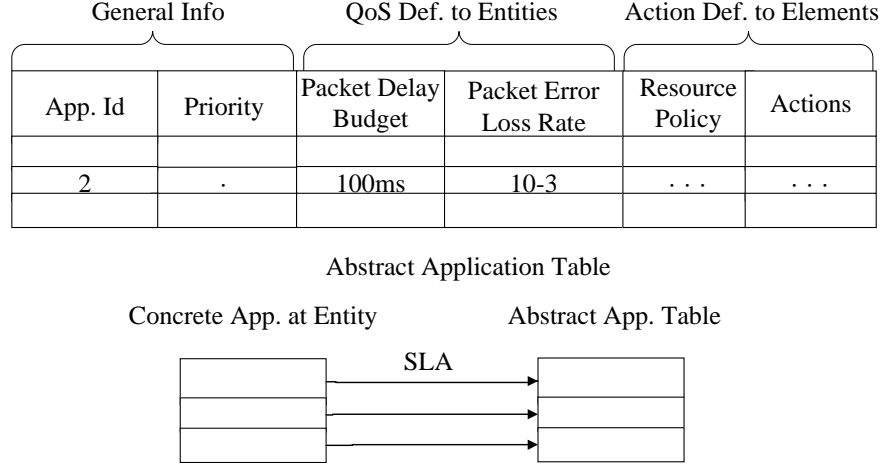


Fig. 3. Definition of Abstract Application and SLA Model.

(resource policy, network actions, *etc.*), and (possibly) unit pricing information. An SLA then indicates how to map concrete applications to abstract applications and the bandwidth demand upon each abstract application. SOARAN thus adapts differentiated services [9] supported at respective entities to the abstract application set. In this way, SOARAN is able to react quickly to fast-growing emerging applications, which will become the new norm in future networks [10], by adding entries at the abstract application table. SOARAN also eases network management and resource allocation by abstracting numerous external nonuniform services in a controllable set. As a result, entities are only required to determine the types and the bandwidths of abstract applications they need on a more trackable and readable pricing system produced by the re-modeled SLA.

According to the bandwidth requirements of respective abstract applications gathered from SLAs, the SOARAN controller configures a lower bound and an upper bound of resource available to each application to enforce application-level isolation, while the resource within the bounds is adjustable and periodically allocated to each application at the time order of several seconds to promote resource utility with respect to dynamic traffic demands and varying wireless channel conditions.

B. The SOARAN Software Architecture

The software architecture of SOARAN is illustrated in Fig. 4. Flows of concrete applications in entities are mapped to abstract applications upon entering the RAN gateway according to respective SLAs. Inside the SOARAN gateway, per time period τ (in the order of seconds), the controller estimates the bandwidth

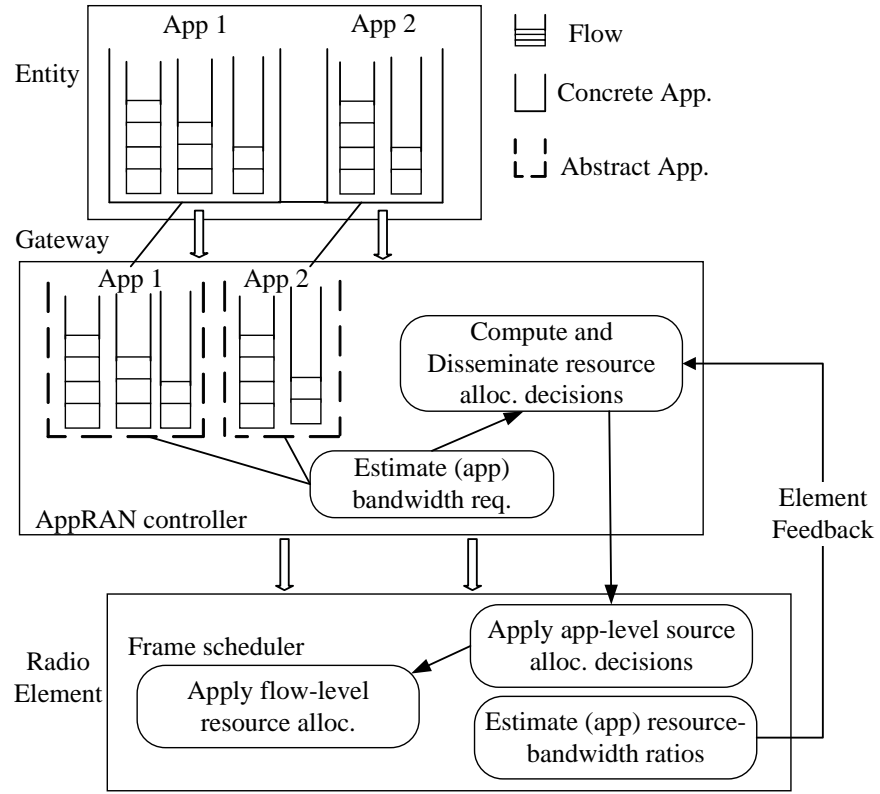


Fig. 4. Software Architecture of SOARAN.

requirement of each application (possibly) through history analysis. On the same periodic scale, each radio element estimates the average resource-bandwidth ratio associated with each application, *i.e.*, average resource per unit data rate, capturing the average channel condition of corresponding application, and sends these ratios to the controller as element feedback. The rationale behind this is that the resource to support unit data rate is jointly determined by QoE requirements (indicated by types of applications) and wireless channel conditions.

Together with the estimated bandwidth requirements, the controller computes the dynamic resource requirement of each application and determines how the resource should be divided among these applications. The calculated resource allocation decisions are then disseminated to respective elements as resource *policies*. Rather than specifying resource allocation for each flow, SOARAN creates resource policies at the application level while granting the flexibility to elements on how to accomplish these policies. We comment that this design is in line with the principle of the SDN architecture [11]. As the radio element has more accurate information of channel conditions as well as fine-grained power management and maturing modulation and coding technologies [12], allowing elements to optimize local

resource allocation will further improve resource utility and enable heterogeneous network deployment [12].

III. OPTIMAL RESOURCE ALLOCATION IN SOARAN

The resource allocation scheduler is the kernel of SOARAN, running at the logically central controller (see Fig. 4), which periodically computes the resource distribution among the applications with respect to current network conditions and system configurations.

A. The Optimization Framework

1) *Formulation:* We target a RAN configuration with a set of I radio elements supporting a set of K applications. Each radio element $i \in I$ has a resource amount B_i , representing the available radio resource at the element, abstracted from the three-dimensional resource grid. The aggregate radio resource over all elements is denoted by B , i.e., $B = \sum_{i \in I} B_i$. For isolation purpose, each application $k \in K$ reserves a minimum resource of L^k irrespective to traffic demands, while it can use up to M^k resource to improve its performance, where $L^k \leq M^k \leq B$. Similarly, the system also configures a lower bound l_i^k and an upper bound m_i^k for resource allocation to application k at element i to enable element-level isolation, accordingly, $L^k = \sum_{i \in I} l_i^k$ and $M^k = \sum_{i \in I} m_i^k$.

Let s_i^k be the amount of resource allocated to application k at element i , $i \in I, k \in K$. Now for each time period τ , we aim to maximize the overall resource allocation gain or utility while assuring that the resource used by each application is bounded according to preset configurations. Defining the utility function $u_i^k(\cdot)$, we formulate the following optimization problem:

$$\begin{aligned}
 & \max \sum_{k \in K} \sum_{i \in I} u_i^k(s_i^k) \\
 & \text{s. t. } \sum_{k \in K} s_i^k \leq B_i, \quad \forall i \in I \\
 & \quad \sum_{i \in I} s_i^k \leq M^k, \quad \forall k \in K \\
 & \quad \sum_{i \in I} s_i^k \geq L^k, \quad \forall k \in K \\
 & \text{var. } 0 \leq l_i^k \leq s_i^k \leq m_i^k, \quad \forall i \in I, k \in K.
 \end{aligned} \tag{1}$$

The first constraint indicates the resource limit at each radio element i . The second and third constraints impose the upper bound and lower bound of resource that can be allocated to each application k over the RAN, respectively. The per-element resource upper bound and lower bound for each application are imposed by the last constraint.

A similar optimization problem is formulated in [6], which attempts to maximize the aggregate utility of allocating network-wide radio resource proportionally to entities or mobile virtual network operators. However, they only consider the symmetric scenario by assuming that all base-stations (termed as radio elements in this paper) possess the same amount of resource, which prevents their framework from scaling to the complicated networking reality nowadays with heterogeneous wireless elements. On the contrary, our formulation directly addresses the amount of radio resource, allowing heterogeneity of elements. Providing the information of available resource at each element, these resource amounts are readily to be converted to percentages for implementation purposes. Therefore, problem (1) fundamentally differs from the model formulated in [6].

2) *Utility Function and Demand Estimation:* The utility function $u_i^k(\cdot)$ can be a linear function or any concave function following the law of diminishing marginal utility, representing the utility value of allocated resource to application k at element i . The following equations show examples of a linear function and a logarithmic function drawn from the proportional fairness principles defined in [13]:

$$\begin{aligned} u_i^k(s_i^k) &= w_i^k \cdot d_i^k \cdot s_i^k, & \text{or} \\ u_i^k(s_i^k) &= w_i^k \cdot d_i^k \cdot \log(s_i^k), \end{aligned} \quad (2)$$

where w_i^k is the utility weight, d_i^k is the resource demand for application k at element i in current period.

In conventional RAN sharing models, it is intractable to estimate resource demand d_i^k with respect to distinct entity, let alone to support differentiated QoE for different applications within one entity. For one thing, the network-level resource allocator has no information of the ownership of flows, *i.e.*, to which entity each flow belongs. This forces the system either to use the off-line, long-term estimation of “average” demands, or to employ deep packet inspection (DPI) to extract application-level information from flows. The former lacks accuracy, while the latter apparently introduces an intolerable computational overhead. For another, translating flow-level bandwidth demands to radio resource demands requires the

information of modulation coding schemes (MCSs) selected for each flow transmission, which in turn relies on element-level details of users' channel conditions and MCS adaptation schemes [12].

Bandwidth demands of ongoing flows in SOARAN are irrespective of the entities they belong to, requiring no DPI operations. Mobile systems usually support Differentiated Services (DiffServ) [9] in their IP backbones for QoE management, *e.g.*, evolved packet system (EPS) bearers in LTE systems [7]. With DiffServ, SOARAN easily determines to which application a flow belongs by checking the QoE class identifier (QCI) attached to the flow. Moreover, together with the knowledge of channel conditions and scheduling algorithms, each radio element is ready to translate bandwidth demands of any application to resource demands. We define the average bandwidth-resource translating ratio for application k at radio element i as $p_i^k = \frac{\text{resource to support application } k \text{ with QoE}}{\text{bandwidth demand for application } k}$, which is reported to the central controller for resource-demand estimation. Here, we note that such translating ratio might not reveal the “real” relation between bandwidth and resource demands in the cases with significant flow fluctuation, *e.g.*, the traffic demand of the user with the worst channel condition soars for the next time period τ , or the channel condition of a heavy-traffic user significantly changes. Yet we argue that our approach remains effective. This is because: (1) For a short time period of τ , it is less likely to have large fluctuation. Even with large fluctuation, the system only experiences suboptimal resource allocation for at most τ time; and (2) In SOARAN, we compute the resources allocated at each element to distinct applications. Therefore, the fluctuation can be mitigated or shaped by employing adaptive MCS schemes [14] at radio elements. That is, given an application and allocated resource, the element runs a second-phase resource allocation to distribute resource among flows with accurate channel state information.

3) *Problem Hardness*: The utility function is either a linear function or a concave function as shown in (2), resulting in a linear programming (LP) model and a nonlinear programming (NLP) model for problem (1) respectively. However, in both models, problem (1) has a prohibitively large size for a direct solution from state-of-the-art LP/NLP solvers, *e.g.*, CPLEX, OPT++. To have a rough understanding, a production mobile system usually has $O(10^5)$ radio elements and supports $O(10^2)$ applications. Therefore, the rudimentary size of problem (1) is with $O(10^7)$ variables and $O(10^7)$ constraints (see the last constraint of problem (1)). Therefore, a fast algorithm with approximate guarantees is more desirable.

B. The Approximate Algorithm

1) *Main Procedure:* We employ the Barrier method from [15] and solve problem (1) via an interior-point approach. For ease of presentation, let \mathbf{s} be the vector of variables $\{s_i^k | i \in I, k \in K\}$. We define the logarithmic barrier function as

$$\begin{aligned} \phi(\mathbf{s}) = & \sum_{i \in I} \log(B_i - \sum_{k \in K} s_i^k) + \sum_{k \in K} \log(M^k - \sum_{i \in I} s_i^k) \\ & + \sum_{k \in K} \log(\sum_{i \in I} s_i^k - L^k). \end{aligned} \quad (3)$$

We denote the objective of problem (1) by $u(\mathbf{s}) = \sum_{k \in K} \sum_{i \in I} u_i^k(s_i^k)$. We then introduce a multiplier t and consider the following problem:

$$\begin{aligned} \max \quad & t \cdot u(\mathbf{s}) + \phi(\mathbf{s}) \\ \text{s. t.} \quad & l_i^k \leq s_i^k \leq m_i^k, \quad \forall i \in I, k \in K. \end{aligned} \quad (4)$$

The main procedure of our algorithm is listed in Algorithm 1. The procedure follows a typical route of the Barrier method, while we develop a tighter bound. Starting from a feasible point, it iteratively solves a sequence of problem (4) with increasing t till $t \geq (B + |K|)/\epsilon$ (to be discussed later). For simplicity, we set $\mathbf{s}_0 = \mathbf{l}$ to be the initial starting point, providing that \mathbf{l} is a feasible solution to problem (1) under proper configurations, *i.e.*, $\sum_k l_i^k \leq B_i, \forall i \in I$. Line 4 therein is called an inner loop for solving the optimization problem (4) with bounded variables. We refer readers to [15] for the details of the inner loop, *e.g.*, the Newton's method. Each solution found in line 4 is then used as a new starting point for the next iteration in the outer loop. Here, μ is a parameter involving a trade-off in the number of iterations of the inner and outer loops. Details on selecting μ can also be found in [15].

Algorithm 1: Barrier method for problem (1).

- 1: Start with an interior feasible point \mathbf{s}_0 ;
 - 2: **while** $(B + |K|)/t > \epsilon$ ($\epsilon > 0$) **do**
 - 3: $\mathbf{s} := \mathbf{s}_0$, $t := t_0$, where $t_0 > 0$;
 - 4: With starting point \mathbf{x} , solve (4) via a gradient-based method and output the solution \mathbf{x}^* .
 - 5: $\mathbf{s} := \mathbf{s}^*$, $t := t \cdot \mu$, where $\mu > 1$;
 - 6: **end while**
-

2) *The Approximate Result:* We now develop the theoretical basis for Algorithm 1. Applying the duality analysis, we have the following conclusions.

Lemma 1. *If problem (4) can be optimally solved, then we can find a solution to problem (1) that is at most ϵ -suboptimal, for any $\epsilon > 0$. In other words, let u^* be the optimal value of problem (1) and \bar{s} be the optimal solution to problem (4). By setting $t \geq (B + |K|)/\epsilon$, we have*

$$u^* \leq u(\bar{s}) + \epsilon.$$

Theorem 1. *For any $\epsilon > 0$, Algorithm 1 obtains a solution to problem (1), which is at most ϵ -suboptimal.*

IV. SIMULATION AND NUMERICAL RESULTS

In this section, we study the SOARAN model through extensive computer simulations. We focus on the application-level resource allocation at the RAN gateway, while we assume that certain flow-level resource allocation schemes, *e.g.*, [12], are adopted by each radio element and radio resource therein is abstracted and represented as a non-negative real value using the technologies in [4,16].

A. Simulation Setup

We simulate a RAN system with 1,000 radio elements shared by 20 entities. SOARAN defines 100 abstract applications with resource requirement factors uniformly selected within $[0.1, 2]$ per unit data rate (Mbps), representing respective QoE guarantees. We assume that all these applications are supported over all entities and radio elements to exclude the complexity of SLAs from our simulations. Given that the information of channel conditions is available to elements, the average resource-bandwidth multipliers are uniformly generated from $[1.0, 2.0]$, resulting in a resource-bandwidth ratio range of $[0.1, 4.0]$ (jointly determined by channel conditions and QoE requirements). The available resource at elements is abstracted as real values randomly generated from $[100.0, 300.0]$ with the mean value of 200.0 over all elements. This setting is to ensure that the logarithmic utility function results in a non-negative value in all cases, representing proportional resource capacities that can be arbitrarily scaled up/down over the system. Two most data-consuming applications, m and n , *e.g.*, video streaming and FTP file downloading, can use at least 5% and up to 40% of the aggregate resource, while other applications equally share the rest resource.

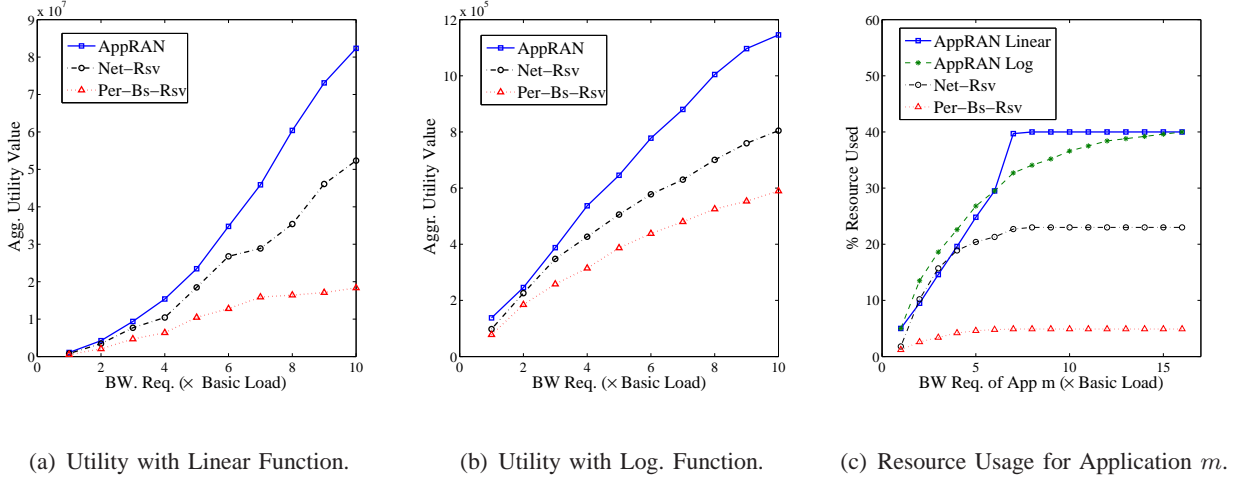


Fig. 5. Performance comparison of SOARAN and operator-oriented alternatives.

For comparison, we align SOARAN with two alternative operator-oriented resource allocation schemes, termed as *Per-Base-station Reservation* (Per-Bs-Rsv) [5] and *Network Reservation* (Net-Rsv) [6], in which the utility is calculated over flows instead of entities as in the literature. In Net-Rsv, each entity reserves 2% of the aggregate resource and can use up to 10%, while in PerBs-Rsv, each entity reserves 5% resource at each radio element. For comparison fairness, the utility is calculated over flows for all three schemes to be comparable, since the accumulated utility either over applications or entities can be decomposed in the form of flows. The basic system load contains 5,000 flows randomly generated from all applications across all radio elements, the bandwidth demands of which are selected from $0.1 \sim 1$ Mbps to ensure that a feasible resource assignment can be reached by all schemes. This load is then gradually increased by a controlled multiplier for different scenarios.

B. Utility Results

For simplicity, we set unit utility weights $w_i^k = 1 \forall i \in I, k \in K$. We use the linear function and the logarithmic function (2) for utility calculations and show results in Fig. 5(a) and Fig. 5(b), respectively. In both scenarios, the system load increases step-by-step from 1 to 10 times of the basic load. With a linear utility function, the quadratic-like utility-growing curve of SOARAN in Fig. 5(a) shows that SOARAN tends to allocate resource linearly to corresponding bandwidth demands when the system is under a low to moderate load. This growth is flattened with a logarithmic utility function (Fig. 5(b)), which also considers the fairness among applications. In both cases, SOARAN obtains a utility objective which significantly

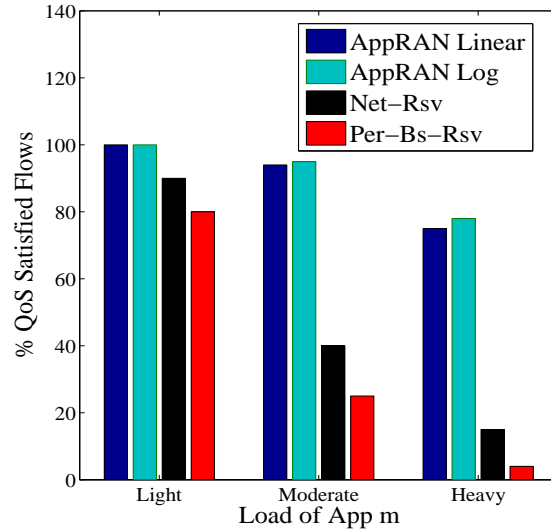


Fig. 6. Comparisons of QoS Satisfied Flows.

outperforms the Net-Rsv scheme (up to 40%) and the Per-Bs-Rsv scheme (up to 200%). This confirms the conclusion that resource reservation over entities immensely limits the RAN sharing performance with multiple QoE-differentiated applications.

C. QoE Results

In this simulation, we study the QoE performance of different schemes by gradually increasing the bandwidth demands of application m , one of the most data-consuming applications. Beyond the basic load setup in Section IV-A, we add 2,000 extra flows of application m from 5 entities to randomly selected 200 radio elements with a mean basic bandwidth demand of 1.0 Mbps. The load of application m is then iteratively increased from 1 to 15 times of the basic load, while the demands of rest flows keep unchanged. Fig. 5(c) shows the resource consumption of application m . It indicates that SOARAN effectively adapts resource allocation for data-consuming applications as the traffic demands increase. In contrast, as constrained by per-entity resource limits, both Net-Rsv and Per-Bs-Rsv result in significant resource under-utilization irrespective of idle resource in the system. In SOARAN, we also observe that with the linear utility function, the resource usage grows more aggressively to its resource upper bound.

In Fig. 6, we show the number of QoE satisfied flows in each case. Here, light, moderate and heavy loads correspond to 1, 10 and 15 times of the basic load of application m . We can see from Fig. 6 that

SOARAN achieves similar QoE performance with the linear or logarithmic utility function. However, with entity-oriented resource reservation, Net-Rsv cannot even fully support all flows with the least load and the performance deteriorates further as the load increases. Likewise, Per-Bs-Rsv produces the worst performance due to its strict entity-oriented resource constraints.

V. CONCLUSION

In this paper, we propose SOARAN, an service-oriented framework for sharing RAN resource in evolving mobile networks. SOARAN encapsulates the complexity of underlying resource management by defining an application abstraction layer for entities and provides QoE-guaranteed services to them. SOARAN centrally optimizes resource distribution among applications at each element, while the decisions on allocating resource to flows are determined distributively at each element with real-time channel conditions. By decoupling RAN sharing participants from radio resource, SOARAN is in line with the principles of SDN and enables better network abstraction. A fast algorithm has been proposed and studied for service-level resource allocation. The simulation results demonstrate significant performance improvement of SOARAN over entity-oriented schemes in terms of aggregate utility and QoE satisfaction.

REFERENCES

- [1] "GSMA report: Mobile infrastructure sharing," <http://www.gsma.com/publicpolicy/mobile-infrastructure-sharing-report/>, November 2008.
- [2] Telcordia, "With the right support, MVNOs can enrich network operators with innovation, differentiation, and market share," White Paper, 2012.
- [3] M. Hoffmann and M. Staufer, "Network virtualization for future mobile networks: General architecture and applications," in *IEEE International Conference on Communications Workshops (ICC)*, 2011, pp. 1–5.
- [4] S. Katti and L. E. Li, "Radiovisor: A slicing plane for radio access networks," *Open Networking Summit 2014 (ONS 2014)*, 2014.
- [5] R. Kikku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, 2012.
- [6] R. Mahindra, A. Khojastepour, H. Zhang, and S. Rangarajan, "Network-wide radio access network sharing in cellular networks," in *Proceedings of IEEE International Conference on Network Protocols (ICNP)*, 2013.
- [7] S. Sesia, I. Toufik, and M. Baker, *LTE: The UMTS Long Term Evolution*. Wiley Online Library, 2009.
- [8] J. He and W. Song, "Appran: Application-oriented radio access network sharing in mobile networks," in *IEEE International Conference on Communications (ICC)*, 2015, pp. 3788–3794.
- [9] K. Nichols *et al.*, "RFC 2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," 1998.

- [10] ADVA *et al.*, “Horizon 2020 advanced 5G network infrastructure for future Internet PPP, draft version 2.1,” 2013.
- [11] B. Nunes, M. Mendonca, X. Nguyen, K. Obraczka, and T. Turletti, “A survey of software-defined networking: Past, present, and future of programmable networks,” 2014.
- [12] Z. Han and K. R. Liu, *Resource Allocation for Wireless Networks*. Cambridge University Press, 2008.
- [13] G. Tychogiorgos, A. Gkelias, and K. K. Leung, “Utility-proportional fairness in wireless networks,” in *IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2012, pp. 839–844.
- [14] S. T. Chung and A. J. Goldsmith, “Degrees of freedom in adaptive modulation: A unified view,” *IEEE Transactions on Communications*, vol. 49, no. 9, pp. 1561–1571, 2001.
- [15] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2009.
- [16] A. Gudipati, D. Perry, L. E. Li, and S. Katti, “SoftRAN: Software defined radio access network,” in *Proceedings of 2nd ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, 2013, pp. 25–30.